Deep Neural Network-empowered Polygenic Disease Prediction on Cardiovascular Diseases

Zelia Soo
Australian Artificial Intelligence Institute
University of Technology Sydney
Ultimo, Australia
zeliamei-ze.soo@student.uts.edu.au

Hua Lin
23Strands
Pyrmont, Australia
hua.lin@23strands.com

Yue Yang
Australian Artificial Intelligence Institute
University of Technology Sydney
Ultimo, Australia
yue.yang-4@student.uts.edu.au

Mark Grosser
23Strands
Pyrmont, Australia
hua.lin@23strands.com

Yi Zhang
Australian Artificial Intelligence Institute
University of Technology Sydney
Ultimo, Australia
yi.zhang@uts.edu.au

Jie Lu
Australian Artificial Intelligence Institute
University of Technology Sydney
Ultimo, Australia
jie.lu@uts.edu.au

***Abstract*— There is a growing area of research showing that complex diseases have been found to be caused by significant genetic variants, that is, multiple changes to the normal genome across multiple locations. Predicting the risk of these diseases is difficult due to the limited knowledge of variant causation and the leading approaches currently focus on gene-disease association. In this work, we propose a cardiovascular disease based analysis using an enhanced indel deep neural network (EI-DNN), comprised of two deep neural networks using novel indel variants alongside conventional variant sites to predict disease risk. This model uses two deep neural networks in series, the first to process indel data and the second to provide the risk score. The experiments were performed on our proposed algorithm using the MGRB database and compared against a conventional PRS calculation and a single DNN algorithm. The experimental results validate the effectiveness of the proposed method and highlight the capabilities with combining indel variants to enhance disease prediction.**

Keywords* — *deep neural network, CVD, genomics, bioinformatics, disease prediction

I. Introduction

Genomics is the study of how your DNA reproduces the cells in your body. Genetic diseases are diseases caused by changes in an individual's genome. These conditions can be categorized into monogenic, denoting diseases that are caused by one genetic variant, and polygenic, denoting diseases that are caused by several variants across the genome that have some contribution to the disease risk [1]. Common diseases are often most likely polygenic [2]. This makes polygenic diseases much more complicated to understand. This is further exacerbated since there are two sets of each chromosome. If there is one variant on one chromosome but not the other this is called heterozygous, and if both chromosomes have the same variant this is called homozygous. Thus, polygenic diseases can vary based on several variants and the severity of the disease could be affected whether an individual is heterozygous or homozygous for all these different variants [3].

Cardiovascular diseases (CVD) are a group of a highly studied polygenic diseases [4]. It is the largest cause of death globally, with 10.8 million deaths attributed to cardiovascular disease and contributed to 11.3 million deaths in 2021 [5]. To help in prevention of CVD, many studies have been undertaken to understand the many gene-disease associations and identify people with high risk of future CVD [4]. However, the accurate prediction of CVD continues to remains a challenge [4].

The relationship of genetic variants and disease has been difficult to fully understand, particularly as the detection of variants has outpaced the ability to understand the relationship between variants and pathogenic outcomes [6]. Efforts to understand the gene-disease relationship have been undertaken in the field of bioinformatics, as seen in many gene-disease or variant-disease association algorithms such as Bibliometric Engine or databases such as ClinVar or Human Phenotype Ontology [7]. Since several genes are associated with polygenic diseases, it is commonly assessed using a polygenic risk score (PRS) also known as a polygenic score (PGS) [8]. Many genetic diseases and traits have PRS information available including CVD such as coronary artery disease and hypertension, multiple cancers such as breast cancer and pancreatic cancer, and lifestyle diseases such as type 2 diabetes [8]. Publicly available catalogues such as the PGS catalogue and Genome-wide association studies (GWAS) catalogue have collated numerous associations from GWAS and other papers to implement PRS findings onto new data [9].

PRS uses a linear regression algorithm to provide a disease risk score for each individual patient [10]. This algorithm is constructed using all trait-associated genetic variants and their associated weightings, as seen in equation 1, where N is the number of variants, β_i is the effect size, and X_i is the genotype or variant.

$$PRS = \sum_{i=0}^{N} \beta_i X_i \qquad (1)$$

While this method has been popular to implement, there are several associated limitations. The only genetic variants that are used in constructing a PRS are single nucleotide polymorphisms (SNPs), a variant in which a single base pair is swapped for another [10]. This excludes many different types of variants such as structural variants and small insertions or deletions (indels) which could be key contributions for a PRS. Even within the SNP group, only common SNPs are utilized. This is defined as a called as minor allele frequency of > 0.01, meaning at least 1% of the global population has this variant [10]. Furthermore, the data cleaning steps are very aggressive and many SNPs are removed in this process [10]. This minute selection of chosen variants limits the scope of an individual's whole genomic data, oversimplifying what variation is investigated within a PRS [4]. Another limitation is the weightings provided for each SNP in a PRS calculation catalogue. GWAS have been undertaken to investigate several polygenic traits and define associated SNPs and SNP weightings based on its p-values and odds ratios [11]. However, there is no distinction between whether this weight only applies to variants that are homozygous or heterozygous, or whether the weighting changes depending on this state. Another limitation is the equation PRS is based on. The simplistic linear nature of equation 1 means the aforementioned limitations do not have a good way of being integrated into the risk score calculation.

Using this PRS methodology tends to produce prediction scores with low accuracy. Thus, conventional PRS have not been able to be used clinically for several traits and diseases [1]. To improve the accuracy of PRS and make it a useful tool in bioinformatics, there have been efforts to utilize machine learning (ML). This is mainly because the linear nature of the original PRS framework is limiting, and ML allows many features or variants to be investigated [12]. Several classical ML algorithms have been tested due to the ease of use from accessible packages in popular programming languages, such as scikit-learn in python [13]. Consequently, more complex algorithms have also been studied, with deep neural networks (DNN) being a popular ML algorithm to test [14]. However, there are still limitations with these DNN approaches, namely still using only GWAS SNP data which restricts the use of a larger use of genomic data.

This paper presents an enhanced indel DNN (EI-DNN), comprised of a double DNN architecture that incorporate other structural variants such as indels alongside SNPs to generate a more accurate and precise PRS. To validate this proposed algorithm, we compared the results of the conventional PRS calculation with our EI-DNN architecture using a cohort from the Medical Genome Reference Bank (MGRB). Our results suggest that deep learning can be used to predict polygenic traits more accurately as well as identify other necessary genetic information that can enhance the predictive ability of PRS.

In this paper, our main contributions are as follows:

- We developed EI-DNN to integrate critical genomic data into PRS calculation and compared it against a single DNN and conventional PRS algorithms to demonstrate the advantage of our approach.
- By incorporating structure variants into the risk score calculation, our model offers an example of an intelligent bioinformatics approach to PRS.

The structure of this paper is as follows: Section II presents the related work. Section III summarizes the methodology. Section IV outlines the experiment undertaken to test the methodology. Section V presents and discusses the results from the experiments. The last section offers conclusions and directions for further research.

II. Related Work

To establish our EI-DNN method, the following areas need to be studied: DNN, combining machine-learning techniques, and genetic-based disease prediction.

A. Deep Neural Networks

DNN have been a popular ML algorithm implemented in several fields due to its flexibility, ability to handle large complex datasets, and learn non-linear relationships [15]. In the field of genomics and bioinformatics, DNNs have been used for many applications as well [13]. Yang et al. used a DNN to predict gene-disease relationships for Parkinson's disease [16]. They combined multi-view phenotype features with genotype features as input into the DNN and the resulting output was a vector that represented the disease and genes. The experiment had a precision improvement of 9.55% and recall improvement of 9.63% compared to other standard algorithms. Another use of DNN was for clustering and dimension reduction of RNA data, a similar type of dataset to DNA. Peng et al. used the Gene Ontology database into the DNN to reduce features, cluster data, and ultimately identify different cell types [17].

In PRS applications, DNNs have been previously tested to compare against conventional PRS and other ML algorithms. Badre et al. tested several ML algorithms and found that a 5-layer DNN model had the best accuracy for disease prediction of breast cancer [12]. Similarly, Zhou et al. found that a 7-layer DNN was more accurate for Alzheimer's disease prediction over PRS algorithms [14].

B. Ensemble Neural Networks

Ensemble methods are also a very popular approach to machine learning. The performance of these ensemble methods can improve by combining predictions from multiple models [15]. In the case of neural networks, there are several that operate using an integration of multiple network types in series or in parallel. Some examples of these include generative adversarial networks for generative models and Siamese networks for discriminative models. Cheng et al. used a GAN model in gene classification and variant detection [18]. A generative model was used to alter the genetic sequence while the discriminative model would compare the genetic sequence and identify what genes were altered [18]. The success of this model demonstrates the power of processing unseen genetic data with unknown patterns and the capabilities of the model to adapt to a biological understanding and explain the problematic nature. Koh et al. used a Siamese network to capture the similarities between different cells and transfer the labelled cell annotations from a single cell RNA dataset to an unannotated cell [19]. This network was able to flag novel cells not in the original dataset and cluster these into new subtypes as well.

C. *Genetic-based Disease Prediction*

Several PRS algorithms have been developed based on the basic linear regression improve the accuracy of these scores. PLINK is an open-source C/C++ software first developed in 2007 by Purcell et al. that is extensively used in a wide variety of genetic analyses [20]. While not solely for the use of PRS, it is able to perform all the steps in calculating PRS using the conventional clumping and thresholding method (C+T), where SNPs are filtered to be independent from each other, removing linkage disequilibrium (LD), so the effects can be summed together. This is a highly manual process, so Euesden et al. developed PRSice and its successor PRSice2, which follow the same C+T methods, automating the steps in a PLINK pipeline in R, including finding the most predictive p-value threshold for PRS calculation [21]. To improve and utilize the typically removed SNPs that are correlated to each other during PRS, LDpred by Vilhjámsson et al. and its successor LDpred-2 by Privé et al. are methods that utilize Bayesian statistics to improve the PRS accuracy [22, 23]. The PRS is calculated similarly but considers LD SNPs by incorporating an independent reference panel from the same population. Lassosum is a method constructed by Shin Heng Mak et al. that uses LASSO regression to select the most relevant SNPs for PRS and estimates the regression coefficients of the chosen SNPs as the effect weight. It also integrates the LD reference panel and altogether generates a PRS [24].

Indels, a compound abbreviation meaning insertions and deletions are a type of structural genetic variant that are prevalent throughout the genome, found in coding genes, non-coding genes, and regulatory areas [25]. Healthy individuals have indels present in their genome with no associated disease phenotypes, however, there are several disorders that can arise from the presence of indels within a gene. Curtis investigated the use of logistic regression for PRS in the disease case of schizophrenia [26]. To increase accuracy of his score, he added rare and schizophrenia-associated copy number variants, another type of structural genetic variant with the weighted SNPs to add another layer of information to each patient.

III. Methodology

A. *Problem Definition*

While PRS is simple to compute, the predictive power tends to be low, thus making it unsuitable for clinical use and is often used as a check to see whether it aligns with other known tests. Its goal is to predict the risk of disease is severely limited [1]. While it is easy to change the linear regression model to a non-linear one to increase predictive power, there are some further issues. In its current form, PRS is difficult to improve because there are limited features used, only using independent SNPs, and neglecting a range of other variant types. Another key issue that needs to be addressed is how to handle the other variant types and undertake feature selection.

B. *Overview of the EI-DNN architecture*

The EI-DNN is a double deep neural network using novel indel variants alongside conventional variant sites to predict disease risk. To achieve our proposed methods, two DNNs were created in series, inspired by the multiple network structures from the related work and their ability to process complex data. The first DNN was for feature selection of relevant indels from the patient data. The second provided the score calculation.

The entire EI-DNN algorithm is summarized in Fig. 1 and can be defined as follows:

$M = f_\theta(x), M \in [0,1]$ where x is the variant feature vector, M is the probability of the sample given x, θ is the set of learning parameters of the neural network f.

$G = g_\phi(y), G \in [0,1]$ where y is the genes feature vector, G is the output score, and ϕ is the set of learning parameters of neural network g.

We define the value function $V(G, M)$ and training of M and G by maximising $V(G, M)$

$$\max_{M,G} V(G,M) = E_{x\sim p(y)}[logG(y)] + E_{x\sim p(x)}\left[logG\big(M(x)\big)\right] \quad (2)$$

Fine tune training consists of backpropagation.

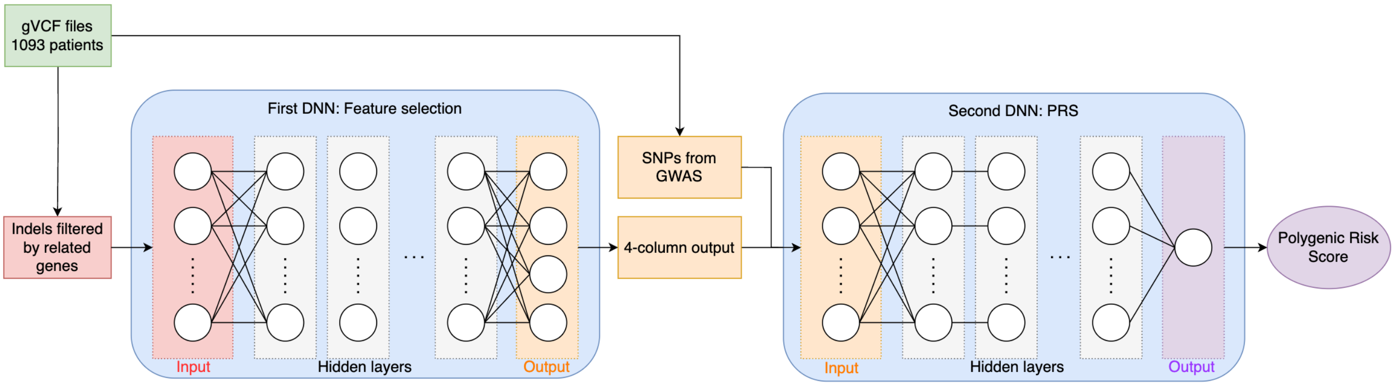

Fig. 1. The EI-DNN framework for polygenic risk score calculation diagram

C. Feature Selection DNN

The input of this DNN was a matrix with all 1093 patients as rows and all discovered indels as features. This DNN constituted of 17 hidden layers: 7 dense layers (activation functions: tanh for the first 3, softmax for the next 2, and sigmoid for the last 2), 5 batch normalization layers, and 5 dropout layers (dropout rate: 0.5 for first, 0.2 for the next 2, 0.1 for the last 2). The output of this DNN is a 4-column matrix which summarizes the polygenic effects of these indels, as Zhou hypotheses that the penultimate nodes may represent different biological process [14].

D. Risk Score Calculation DNN

The input of this DNN was a concatenated matrix from the SNPs from the relevant GWAS catalogue for each patient with the 4-column matrix output from the first DNN. There are 16 hidden layers: 6 dense layers (activation functions: tanh for the first 3, softmax for the next 2, and sigmoid for the last), 5 batch normalization layers, and 5 dropout layers (dropout rate: 0.3 for the first, 0.2 for the second, and 0.1 for the last 3). The output is a risk score denoting a percentage for the binary traits of hypercholesterolemia and hypertension respectively.

IV. Experimentation

In this section, we use the blood pressure and cholesterol labelled data derived from MGRB to evaluate the proposed method. We also conduct comparative experiments on the traditional weighted PRS and the single-DNN algorithm from [14] to demonstrate the predictive performance of our framework.

A. Dataset

The Medical Genome Reference Bank (MGRB) is a dataset funded by the NSW government in Australia [27]. This contains the whole genome sequences (WGS) of 4011 individuals from two studies: the 45 and Up participants from the Sax Institute and ASPirin in Reducing Events in the Elderly (ASPREE) participants from Monash University. This data also contains phenotypic labels for each individual including height, gender, age, and weight. From these participants, 1093 individuals had further information about whether they had treatment for high blood pressure and high cholesterol as a binary "TRUE" or "FALSE". The data from the 1093 participants was cleaned using bcftools, a C-built program that is extensively used in bioinformatics for processing genomic data.

B. Dataset Pre-processing

For indel detection, bcftools and vcftools were used to extract out all indel sites from the entire WGS for all 1093 individuals. From this process, over a million features were identified. To reduce the feature space, a subset was identified using the "chromosome:position" data from relevant genes, an evidence-based natural language processing algorithm [28]. This was then modified to work as input for a neural network by using one hot encoding. Two separate labels for blood pressure and high cholesterol were also added. This is summarized in Fig. 2.

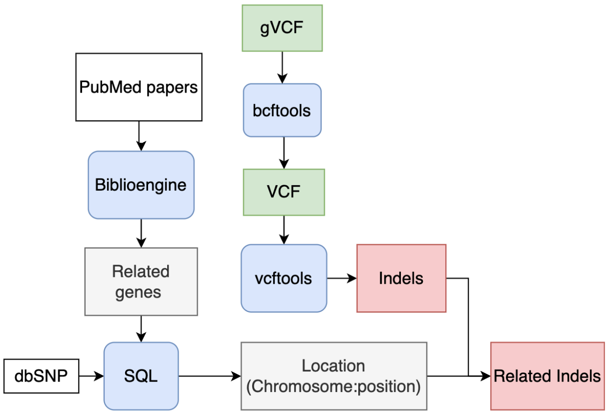

Fig. 2. Flowchart of indel data processing

For SNP detection, PLINK 1.9 was used to extract the SNPs at the locations determined by the relevant GWAS summary statistics from the GWAS catalogue at https://www.ebi.ac.uk/gwas/. The catalogue codes were HP_0003124 for high cholesterol (hypercholesterolemia), and EFO_0000537 for high blood pressure (hypertension). For the weighted PRS experimentation, the PLINK 1.9 files were used. To convert the PLINK 1.9 output into an appropriate input into the neural network, it was converted into a single matrix. This is summarized in Fig. 3.

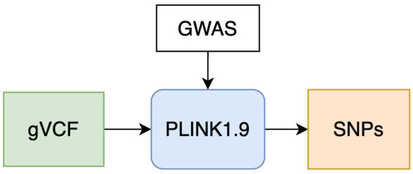

Fig. 3. Flowchart of SNP data processing

C. Standard Baseline

The process of cleaning, sorting SNPs, and final calculation of PRS scores was based on the tutorial by Choi et al. [10], which outlines the quality control steps for the base data from the GWAS and quality control of the target data, obtained for the MGRB. For the base data, the same three GWAS for hypercholesterolemia and hypertension as stated in the SNP variant selection were obtained from the GWAS catalogue. The target data was converted into PLINK readable files and filtered according to Choi et al. The resulting PRS calculated at the end of the process was used as a baseline result for comparison of the EI-DNN algorithm output.

D. Experimental setup

In this study, our central goal is to predict disease risk in individuals using their genome. In our experiments, we had 756 features for hypertension, 12 for hypercholesterolemia from the GWAS association studies respectively, and 306 features for hypertension 327 features for hypercholesterolemia from the indel extraction.

The datasets were split 70% for training and 30% for testing. Furthermore, we used an epoch size of 40 and batch size of 100 for both DNNs.

Our evaluation metrics to determine the strength of each method were the coefficient of determination (R^2), area under the curve (AUROC) and accuracy (Acc.). R^2 is included since it is a conventional method to ascertain the performance of a PRS due to its linearity. A confusion matrix for both traits were also constructed, and the recall, specificity and accuracy were calculated based on equations 3. 4. and 5.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{5}$$

V. Results and Discussion

To evaluate the prediction performance of the proposed EI-DNN method, we examined two diseases and compared to two previous PRS methodologies. Table I gives quantitative results of the weighted PRS, single DNN (sDNN), and EI-DNN, while Fig. 4 and 5 shows the AUROC comparisons. It is found that the EI-DNN achieves the best performance than the other methods for both hypercholesterolemia and hypertension. This verifies that our ensemble DNN framework can improve the predictive power of polygenic risk scores.

Table I.

Comparison of Weighted PRS, sDNN and EI-DNN models

	Hypercholesterolemia			**Hypertension**		
Method	***R²***	***ROC***	***Acc.***	***R²***	***ROC***	***Acc.***
Weighted PRS	0.0192	0.5132	0.5811	0.2011	0.6359	0.5832
sDNN		0.5409	0.6033		0.8939	0.7562
EI-DNN		**0.6961**	**0.6515**		**0.9113**	**0.7945**

In the ROC hypertension case in Fig. 6, sDNN and EI-DNN are much closer together compared to the hypercholesterolemia case in Fig. 5. This is most likely due to the significantly increased feature set from the GWAS database for the hypertension case compared to the hypercholesterolemia case, where the SNPs have been validated to be relevant to this disease.

Furthermore, while the overall accuracy is increased, it appears that when the false positive rate is <0.1, sDNN performs better than EI-DNN in Fig 5. The slightly worse performance at this range may be caused by instability from a smaller sample size.

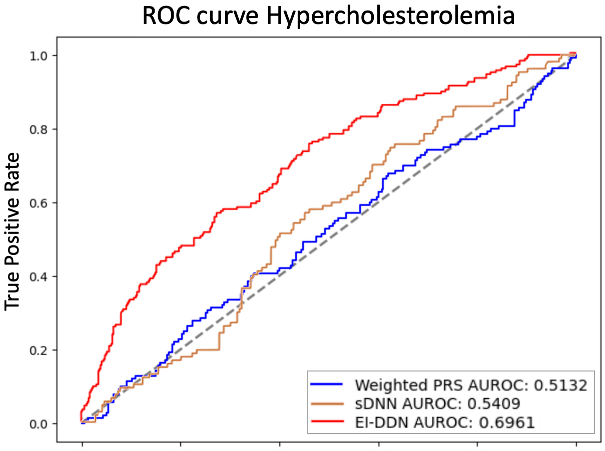

Fig. 4. Comparison of ROC curves for hypercholesterolemia

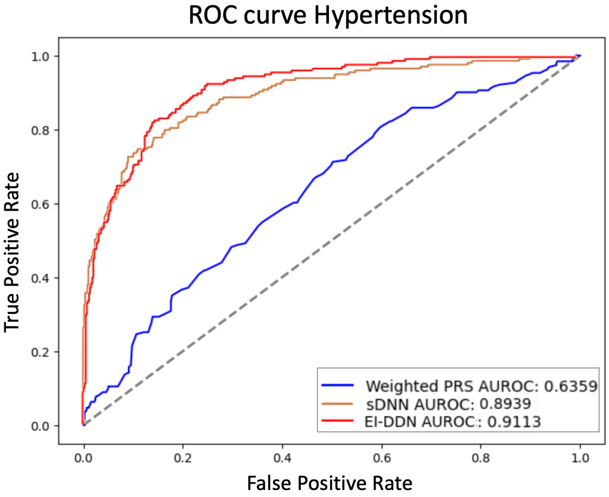

Fig. 5. Comparison of ROC curves for hypercholesterolemia

Table II and III gives the confusion matrices for the cases of hypercholesterolemia and hypertension respectively. Table IV gives a comparison of the recall, specificity, and accuracy of

Table II.

Confusion matrix for hypercholesterolemia

Hypercholesterolemia		**Predicted**	
		Positive	***Negative***
Actual	***Positive***	18	35
	Negative	78	194

Table III

Confusion matrix for hypertension

Hypertension		**Predicted**	
		Positive	***Negative***
Actual	***Positive***	84	44
	Negative	23	175

TABLE IV

RECALL, SPECIFICITY, AND ACCURACY OF PRS, SDNN AND EI-DNN MODELS

	Hypercholesterolemia			**Hypertension**		
	Rec.	***Spec.***	***Acc.***	***Rec.***	***Spec.***	***Acc.***
Weighted PRS	0.3051	0.6332	0.5811	0.3103	0.6344	0.5832
sDNN	0.3433	0.6757	0.6068	0.8224	0.6140	0.7562
EI-DNN	**0.3396**	**0.7132**	**0.6515**	**0.6562**	**0.8838**	**0.7945**

The results in Table IV of the hypertension case were superior to the hypercholesterolemia case, similar to results depicted in the ROC curves in Fig 5. and Fig 6. Since there are only 12 SNPs in the GWAS that were provided, the weighted PRS and sDNN models which rely solely on these have bad accuracy and AUROC scores. Having the indel output seems to provide an increased prediction capability for both, but most notably the hypercholesterolemia test due to the lack of features initially.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed EI-DNN, a double DNN framework for the improvement of polygenic risk score prediction. EI-DNN is trained by making use of a second DNN to process indel data, a type of variants rarely used in normal PRS calculations. Compared with many classical algorithms in recent years, the proposed method has achieved better experimental results and is verified by experiments on real patient data.

Our proposed approach still has limitations. The labels given with this dataset was limited and so only two binary diseases could be tested. Furthermore, this cohort only comprises of relatively healthy individuals, so a large proportion are negative for these traits, and such traits would be minor compared to more severe cases. Further work is now planned for other less studied diseases, such as endometriosis, which has substantially less research compared to CVD. Additionally, expanding to larger databases which has a larger number of patient cases would allow for further validation and could provide opportunity to develop a more robust algorithm. This includes multi-modal databases that have further information than purely genomic. Another area for further work would be interpreting the DNN and creating space for explainable AI, especially in the bioinformatics area.

ACKNOWLEDGMENT

This research is supported by the Australian Research Council Linkage Program: LP210100414.

Thanks to Lucy Chhuo at 23Strands for helping with data pre-processing.

REFERENCES

[1] C. M. Lewis and E. Vassos, "Polygenic risk scores: from research tools to clinical instruments," *Genome Medicine,* vol. 12, no. 1, p. 44, 2020/05/18 2020, doi: 10.1186/s13073-020-00742-5.

[2] N. R. Wray, C. Wijmenga, P. F. Sullivan, J. Yang, and P. M. Visscher, "Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model," (in eng), *Cell,* vol. 173, no. 7, pp. 1573-1580, Jun 14 2018, doi: 10.1016/j.cell.2018.05.051.

[3] J. A. Collister, X. Liu, and L. Clifton, "Calculating Polygenic Risk Scores (PRS) in UK Biobank: A Practical Guide for Epidemiologists," (in English), *Frontiers in Genetics,* Technology and Code vol. 13, 2022-February-18 2022, doi: 10.3389/fgene.2022.818574.

[4] J. W. O'Sullivan *et al.*, "Polygenic Risk Scores for Cardiovascular Disease: A Scientific Statement From the American Heart Association," *Circulation,* vol. 146, no. 8, pp. e93-e118, 2022, doi: doi:10.1161/CIR.0000000000001077.

[5] M. Vaduganathan, G. A. Mensah, J. V. Turco, V. Fuster, and G. A. Roth, "The Global Burden of Cardiovascular Diseases and Risk," *Journal of the American College of Cardiology,* vol. 80, no. 25, pp. 2361-2371, 2022, doi: doi:10.1016/j.jacc.2022.11.005.

[6] D. Grissa, A. Junge, T. I. Oprea, and L. J. Jensen, "Diseases 2.0: a weekly updated database of disease–gene associations from text mining and data integration," *Database,* vol. 2022, 2022, doi: 10.1093/database/baac019.

[7] Y. Zhang *et al.*, "Framework of computational intelligence-enhanced knowledge base construction: methodology and a case of gene-related cardiovascular disease," *International Journal of Computational Intelligence Systems,* 2020.

[8] S. A. Lambert *et al.*, "The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation," *Nature Genetics,* vol. 53, no. 4, pp. 420-425, 2021/04/01 2021, doi: 10.1038/s41588-021-00783-5.

[9] E. Sollis *et al.*, "The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource," (in eng), *Nucleic Acids Res,* vol. 51, no. D1, pp. D977-d985, Jan 6 2023, doi: 10.1093/nar/gkac1010.

[10] S. W. Choi, T. S. Mak, and P. F. O'Reilly, "Tutorial: a guide to performing polygenic risk score analyses," (in eng), *Nat Protoc,* vol. 15, no. 9, pp. 2759-2772, Sep 2020, doi: 10.1038/s41596-020-0353-1.

[11] N. J. Wald and R. Old, "The illusion of polygenic disease risk prediction," *Genetics in Medicine,* vol. 21, no. 8, pp. 1705-1707, 2019/08/01 2019, doi: 10.1038/s41436-018-0418-5.

[12] A. Badré, L. Zhang, W. Muchero, J. C. Reynolds, and C. Pan, "Deep neural network improves the estimation of polygenic risk scores for breast cancer," *Journal of Human Genetics,* vol. 66, no. 4, pp. 359-369, 2021/04/01 2021, doi: 10.1038/s10038-020-00832-7.

[13] K. W. Guo, Mengjia, Wu; Soo, Zelia; Yang, Yue; Zhang, Yi; Zhang, Qian; Lin, Hua; Grosser, Mark; Venter, Deon; Zhang, Guangquan; Lu, Jie, "Artificial intelligence-driven biomedical genomics," unpublished.

[14] X. Zhou *et al.*, "Deep learning-based polygenic risk analysis for Alzheimer's disease prediction," *Communications Medicine,* vol. 3, no. 1, p. 49, 2023/04/06 2023, doi: 10.1038/s43856-023-00269-x.

[15] H. Osipyan, B. I. Edwards, and A. D. Cheok, *Deep Neural Network Applications*. Milton: Taylor & Francis Group, 2022.

[16] K. Yang *et al.*, "PDGNet: Predicting Disease Genes Using a Deep Neural Network With Multi-View Features," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 19, no. 1, pp. 575-584, 2022, doi: 10.1109/TCBB.2020.3002771.

[17] J. Peng, X. Wang, and X. Shang, "Combining gene ontology with deep neural networks to enhance the clustering of single cell RNA-Seq data," *BMC Bioinformatics,* vol. 20, no. 8, p. 284, 2019/06/10 2019, doi: 10.1186/s12859-019-2769-6.

[18] M. Cheng, Y. Li, S. Nazarian, and P. Bogdan, "From rumor to genetic variant detection with explanations: a GAN approach," *Scientific Reports,* vol. 11, no. 1, p. 5861, 2021/03/12 2021, doi: 10.1038/s41598-021-84993-1.

[19] W. Koh and S. Hoon, "MapCell: Learning a comparative cell type distance metric with Siamese neural nets with applications towards cell-types identification across experimental datasets," ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 2019.

[20] S. Purcell *et al.*, "PLINK: a tool set for whole-genome association and population-based linkage analyses," (in eng), *Am J Hum Genet,* vol. 81, no. 3, pp. 559-75, Sep 2007, doi: 10.1086/519795.

[21] J. Euesden, C. M. Lewis, and P. F. O'Reilly, "PRSice: Polygenic Risk Score software," (in eng), *Bioinformatics,* vol. 31, no. 9, pp. 1466-8, May 1 2015, doi: 10.1093/bioinformatics/btu848.

[22] B. J. Vilhjálmsson *et al.*, "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores," (in eng), *Am J Hum Genet,* vol. 97, no. 4, pp. 576-92, Oct 1 2015, doi: 10.1016/j.ajhg.2015.09.001.

[23] F. Privé, J. Arbel, and B. J. Vilhjálmsson, "LDpred2: better, faster, stronger," *Bioinformatics,* vol. 36, no. 22-23, pp. 5424-5431, 2020, doi: 10.1093/bioinformatics/btaa1029.

[24] T. S. H. Mak, R. M. Porsch, S. W. Choi, X. Zhou, and P. C. Sham, "Polygenic scores via penalized regression on summary statistics," (in eng), *Genet Epidemiol,* vol. 41, no. 6, pp. 469-480, Sep 2017, doi: 10.1002/gepi.22050.

[25] J. M. Mullaney, R. E. Mills, W. S. Pittard, and S. E. Devine, "Small insertions and deletions (INDELs) in human genomes," (in eng), *Hum Mol Genet,* vol. 19, no. R2, pp. R131-6, Oct 15 2010, doi: 10.1093/hmg/ddq400.

[26] D. Curtis, "A weighted burden test using logistic regression for integrated analysis of sequence variants, copy number variants and polygenic risk score," *European Journal of Human Genetics,* vol. 27, no. 1, pp. 114-124, 2019/01/01 2019, doi: 10.1038/s41431-018-0272-6.

[27] "Medical Genome Reference Bank." Garvin Institute of Medical Research. https://www.garvan.org.au/research/kinghorn-centre-for-clinical-genomics/research-programs/sydney-genomics-collaborative/mgrb (accessed 24 July, 2023).

[28] M. Wu, Y. Zhang, G. Zhang, and J. Lu, "Exploring the genetic basis of diseases through a heterogeneous bibliometric network: A methodology and case study," *Technological Forecasting and Social Change,* vol. 164, p. 120513, 2021/03/01/ 2021, doi: https://doi.org/10.1016/j.techfore.2020.120513.